ORIGINAL ARTICLE

A comparative analysis to evaluate the effectiveness of Bing, Bard, ChatGPT-3.5 and ChatGPT-4.0 in answering glaucoma-related questions

Mehmet Canleblebici[1], *Ali Dal*[2], *Murat Erdag*[3]

ABSTRACT

Purpose: Large language models can be used for education and training in glaucoma theoretically. The aim of study is to determine the proficiency and differences of chatbots in the field of glaucoma through self-assessment questions

Materials and Methods: The self-assessment questions in the last decade were obtained from the American Academy of Ophthalmology Basic and Clinical Science Course Glaucoma Section books to be used in the study. These questions were asked one by one to ChatGPT-3.5 and 4.0, Bing and Bard respectively. The answers recorded as true and false were analyzed to evaluate the performance of artificial intelligence chatbots. Questions were evaluated in six main categories. In addition to descriptive statistical methods, the Fisher's exact test and Pearson's chi-square test was used to analyze the chatbots both pairwise and together.

Results: ChatGPT-4.0 had the highest correct response rate at 85.10%. Bing had a good accuracy rate of 81.80%. Bard and ChatGPT-3.5 underperformed, at 67.80% and 64.50%, respectively. There was statistical significance when all groups were compared($p<0.05$). In pairwise comparison, there was a statistically significant difference between ChatGPT-4.0 with Bard and ChatGPT-3.5 and between Bing with Bard and ChatGPT-3.5($p<0.05$). No significant difference was observed between ChatGPT-4.0 and Bing, Bard and ChatGPT-3.5($p>0.05$).

Conclusion: ChatGPT-4.0 and Bing showed an impressive correct response rate, while ChatGPT-3.5 and Bard were unfortunately inadequate. ChatGPT-4.0 and Bing have the potential to be used in education and training if care is taken to avoid misinformation, inaccurate results, and bias. Bard has a low response rate but is open to improvement.

Keywords: Large language models, glaucoma, ChatGPT, Bing, Bard.

INTRODUCTION

Technological advancements in the post-millennium era have progressed swiftly, with notable acceleration in recent years, particularly in the realm of artificial intelligence.[1] The utilization of deep learning and large language models (LLMs) in daily activities has commenced due to significant progress in these fields.[2] LLMs are currently being investigated in various domains within the medical industry, including education, diagnosis, patient care and training, and teaching, and their influence is starting to become evident.[3] LLM-supported AI chatbots have particular utility in the field of medicine for educational and training purposes.[4] The chatbot market has made significant progress this year with the introduction of Chat Generative Pre-trained Transformer-4.0 (ChatGPT-4.0), an enhanced iteration of Open AI's ChatGPT-3.5. Microsoft has integrated this ChatGPT-4.0 system into its own AI, Bing chat, while Google has deployed its Bard AI, resulting in a substantial advancement in the chatbot industry. Several publications were published in 2023 assessing the efficacy of these chatbots in medical examinations.[5,6] Particularly in the discipline of ophthalmology, comprehensive assessments of board examinations were of primary importance.[7,8,9,10]

1- Ophthalmology Department, Kayseri State Hospital, Kayseri, Turkiye

2- Ophthalmology Department, Mustafa Kemal University, Hatay, Turkiye

3- Ophthalmology Department, Fırat University, Elazıg, Turkiye

Received: 14.03.2024
Accepted: 23.07.2024
TJ-CEO 2024; 19: 205-210
DOI: 10.37844/TJ-CEO.2024.19.28

Correspondence author:
Mehmet Canleblebici
Email: mehmetcl@hotmail.com

In this study, we aimed to evaluate the performance of ChatGPT-3.5 and 4.0, Bing and Bard, the most popular and widely used LLM-supported chatbots, on glaucoma questions by comparing them with each other.

MATERIALS AND METHODS

A total of 121 study questions were obtained from the self-assessment test of the last decade in the American Academy of Ophthalmology Basic and Clinical Science Course Glaucoma books. There were a total of 133 questions with four major revisions in the last 10 years. Exam questions from the last 10 years were used in order to increase the diversity of questions, but not to be outdated. Since 12 of these questions were exactly the same, 121 questions were used in the study. For subgroup analysis, questions were divided into six main categories: group 1 as epidemiology and genetic questions, group 2 as anatomy questions, group 3 as pathogenesis and pathophysiology questions, group 4 as questions associated with diagnostic tools and procedures, group 5 as treatment management questions and lastly group 6 as clinical case scenarios are defined. The questions were asked individually to ChatGPT-3.5 and ChatGPT-4.0, Bing and Bard in December 2023. There was no prompting or re-asking. If the chatbot did not indicate the correct answer, the answer was recorded as wrong. To avoid subjectivity, questions were asked as they were in the test without difficulty categorization and were not compared with those of human participants. Responses were recorded as correct and incorrect and evaluated by statistical analysis. Subgroups were analyzed individually for subgroup evaluation. This study is exempt from ethics committee approval as the subject of the study does not include living beings.

Statistical Analysis

Statistical program was performed with SPSS for Macintosh Client 25.0 (2016, IBM, Chicago, IL). Descriptive tests and n (%) were used for categorical variables. Kolmogorov-Smirnov test was performed to investigate normal distribution. Pearson's chi-squared was used to analyze nominal independent values together. Fisher's exact test was used to compare groups pairwise within themselves. A p value less than 0.05 was considered statistically significant.

RESULTS

ChatGPT-4.0 had the highest accuracy rate with 85.10% correct answers. Bing had a good success rate of 81.80%. Bard and ChatGPT-3.5 answered the questions correctly with 67.80% and 64.50% respectively. The rate of all four AIs having the same answer to the question was 47.10%. A statistically significant difference was observed when the results of 4 chatbots were compared together using Pearson Chi square test ($p<0.01$).

The Kolmogorov-Smirnov test result showed that there was no normal distribution ($p<0.01$). Therefore, pairwise comparisons were made using the 2-sided Fisher's exact test. According to these results, the acuity of ChatGPT-4.0 was statistically significantly more accurate than ChatGPT-3.5 and Bard ($p<0.05$). And a statistically significant difference was observed between Bing with ChatGPT-3.5 and Bard ($p<0.05$). There was no statistically significant difference between Bard and ChatGPT-3.5 and between ChatGPT-4.0 and Bing ($p>0.05$). Table 1 shows all comparative results.

There were 23 questions in group 1, 12 in group 2, 23 in group 3, 22 each in groups 4 and 5, and 19 in group 6. ChatGPT-4.0 was the most successful chatbot for groups 1, 3, 4, and 6, showing significant differences only with ChatGPT-3.5 in groups 1 and 6. For group 2, Bing was the most successful, with significant differences from the other chatbots. In group 5, Bing was also the top performer, showing significant differences from the other chatbots except ChatGPT-4.0. No statistically significant differences

Table 1: *Statistical pairwise comparison of accuracy rates to AI chatbots with 2-sided Fisher's exact test.*

AI Chatbots		p value	AI Chatbots		p value
ChatGPT-4.0 vs.	ChatGPT-3.5	**p=0,006**	ChatGPT-3.5 vs.	Bing	**p=0,002**
	Bing	p=0,051		Bard	p=0,075
	Bard	**p=0,002**		ChatGPT-4.0	**p=0,006**
AI Chatbots		p value	AI Chatbots		p value
Bing vs.	Bard	**p=0,045**	Bard vs.	ChatGPT-3.5	p=0,075
	ChatGPT-3.5	**p=0,002**		ChatGPT-4	**p=0,002**
	ChatGPT-4	p=0,051		Bing	**p=0,045**

were observed when each chatbot was analyzed across all subgroups (p>0.05). Figure 1 and Table 2 provide detailed results.

DISCUSSION

Ophthalmology is one of the departments most intertwined with technological developments.[11] LLMs, are a class of AI models designed to understand and construct words and sentences just like humans. Software differences, the data set used in the pre-training phase, and access to data on the internet affect the performance of chatbots.[2]

Among the currently most utilized chatbots, ChatGPT-4.0 stands at the forefront of the most popular conversational assistants. It was released in March 2023, marking a significant advancement over its predecessor, version 3.5.[14] ChatGPT-4.0 has limited internet access, but when

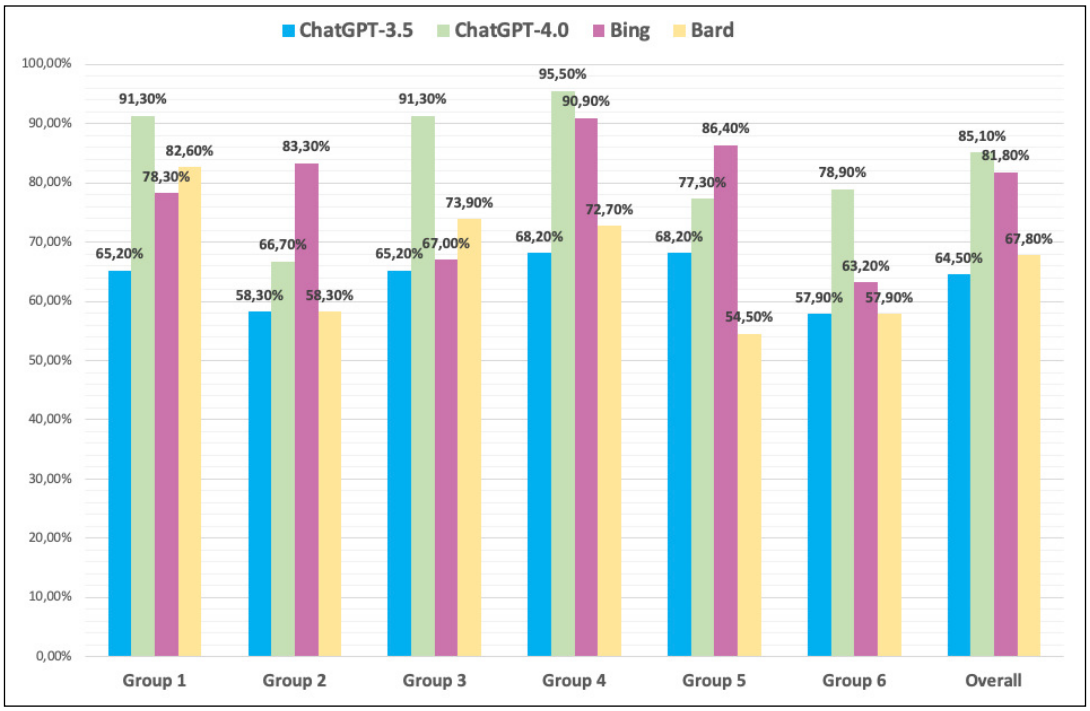

Figure 1: *Accuracy of Chatbots for Glaucoma section self-assessment questions with subgroups. (*Group 1: Epidemiology and genetic questions, Group 2: Anatomy questions, Group 3: Pathogenesis and pathophysiology questions, Group 4: Questions associated with diagnostic tools and procedures, Group 5: treatment management questions, Group 6: Clinical case scenarios.)*

Table 2: *Correct response rates according to Chatbots and statistical comparison of subgroups together.*

	Question numbers	ChatGPT-3.5		ChatGPT-4.0		Bing		Bard		Overall values for each group
		Accuracy	p value	Accuracy	p value	Accuracy	p value	Accuracy	p value	Accuracy
Group 1	23	65,20%	p=0.063	91,30%	**p<0.001**	78,30%	**p=0.018**	82,60%	**p=0.009**	79,35%
Group 2	12	58,30%	p=0.422	66,70%	p=0.286	83,30%	**p=0.004**	58,30%	p=0.448	66,65%
Group 3	23	65,20%	p=0.212	91,30%	**p<0.001**	67,00%	p=0.117	73,90%	**p=0.037**	74,35%
Group 4	22	68,20%	p=0.588	95,50%	**p=0.008**	90,90%	**p=0.033**	72,70%	p=0.302	81,83%
Group 5	22	68,20%	p=0.492	77,30%	**p=0.033**	86,40%	**p=0.001**	54,50%	p=0.451	71,60%
Group 6	19	57,90%	p=0.564	78,90%	**p=0.012**	63,20%	p=0.083	57,90%	p=0.083	64,48%
Overall values for each LLM	121	64,50%	p=0.482	85,10%	**p<0.001**	81,80%	**p=0.006**	67,80%	p=0.368	74,80%

*Group 1: Epidemiology and genetic questions, Group 2: Anatomy questions, Group 3: Pathogenesis and pathophysiology questions, Group 4: Questions associated with diagnostic tools and procedures, Group 5: treatment management questions, Group 6: Clinical case scenarios.
** p values with statically significant are shown with bold characters.
*** The all groups are compared with Pearson Chi Square test.

answering questions, it relies on its own training and data. The lower versions do not possess any features related to internet searches. Both versions are currently available for use (December 2023).

Bing Chat is another LLMs that has been using the ChatGPT-4.0 architecture since April 2023.[15] Microsoft has integrated this chatbot into its Bing search engine. Bing is currently free, has internet access.

Bard is a chatbot that uses the Pathway Language Model (PaLM), which was developed by Google.[16] There is no question limit or hour limit. Bard has internet access. Due to all these differences and diversity, chatbots have different levels of understanding and response. Therefore, evaluating the question-solving capacity of the most widely used LLMs in the field of glaucoma is the main topic of this study.

The question material was obtained from the glaucoma section of the American Academy of Ophthalmology, Basic and Clinical Science Course book, which has been the most fundamental and essential source of current information and techniques for examinations over the years.[17] Sensoy et al. evaluated ChatGPT-3.5, Bing, and Bard over 36 questions using the 2022-2023 Basic and Clinical Science Course Ophthalmic Pathology and Intraocular Tumor Study questions section and found no statistical difference between them for performance. In the study where ChatGPT-4.0 could not be evaluated because it was not released to the market, the question set is similar to our study but covers only one year.[18]

Regarding the overall exam performance of the chatbots, it is seen that ChatGPT-4.0 is the most successful and Bing follows it. While there are various results for Bard and ChatGPT-3.5's performance has started to lag behind compared to the others. In the French version of the European Board of Ophthalmology examination, 6785 questions were asked to ChatGPT-4.0 and 6188 (91.2%) of these questions were answered correctly. For 500 questions in the Japanese board exam, ChatGPT-3.5 and 4.0 were compared with real students and the results were slightly different. While accuracy without prompting was 22% and 45.8%, respectively, v4.0 with less prompting performed 46.2%. Interestingly, v3.5 responses were 2–3 times lower than humans, while v4.0 was close to 70% similar to humans. The main difference may be related to the language being Japanese or the difficulty of the test.[10] In a study in which human participants, ChatGPT-3.5, 4.0 and Bing were evaluated in 250 ophthalmology board style questions from the Basic Science and Clinical Science Self-Assessment Program, their performances were measured at 72.2%, 58.8%, 71.6% and 71.2%, respectively.[8] In another study in which part 1 Fellowship of the Royal College of Ophthalmologists Multiple Choice questions were evaluated in ChatGPT, Bard, and human participants, ChatGPT outperformed humans while Bard was inferior.[9] In a study where ChatGPT-3.5 was evaluated based on 11 questions prepared in glaucoma based on clinical case style, it was correct in 8 cases (72.7%).[19] Our results are similar to those of these studies. By a narrow margin, ChatGPT-4.0 was the most successful. ChatGPT-3.5 lags behind and will probably not improve its performance. However, if Bard can continue its development, it seems to have the potential to increase its success, like other chatbots.

Considering the subgroup analyses, it is observed that although ChatGPT-4.0 is the most superior chatbot on average, Bing is more successful in two subgroups, and Bard is close to ChatGPT-4.0 in two groups. ChatGPT-3.5 shows the lowest correct answer in all subgroups except Group 5. Although Bing shows a statistical difference in the second group with anatomy questions, this section has the least number of questions, a better evaluation could have been made if the number of questions were more. These differences may be due to the fact that the artificial intelligence algorithms mentioned above are different from each other.

Apart from exams, there are studies evaluating chatbots in different areas, like counseling in ophthalmology. Different results were obtained in these studies. ChatGPT was found to be insufficient for emergency eye cases; it was shown that it gave dangerous and inappropriate responses for counseling in vernal keratoconjunctivitis; it was emphasized that it should be more reliable and reproducible in various fields for uveitis; and it was stated that it generally gave appropriate answers about myopia, but it was stated to be careful about inaccurate and misinterpreted answers.[20,21,22,23] Currently, despite their high potential, using LLMs requires caution in areas such as patient counseling and treatment, complication management, and side effect management. Otherwise, the consequences may be to the detriment of the patient.

The fact that chatbots can answer questions in the field of ophthalmology so accurately may allow them to be used in future ophthalmology-based exams in areas such as level determination, knowledge measurement, reliability

of student exams, asking students questions prepared by AI based on student education or accessing the correct answers. These developments may facilitate education in the field of ophthalmology.

Although the easy access of AI to information enables them to provide highly accurate answers to questions on paper thanks to their algorithms, the analysis ability and experience of the human mind make the difference in real life scenarios. In their study, Inayat et al. asked questions, which they created from real on-call pages received within the ophthalmology department at an academic site, to consultant and resident ophthalmologist and ChatGPT.[24] Human participants answered close to 90% correctly, while ChatGPT was around 70%. Importantly, this study indicated that AI has an over-orientation towards emergencies and that human decision-making is important at the tirage stage. In another study investigating ChatGPT's ability to communicate with glaucoma patients, it was stated that ChatGPT was good at providing general information, such as definitions and potential treatment options. However, the readability score and the rate of repeating the same answers were high.[25]

The limitations of this study include the lack of question difficulty classification. However, since the self-assessment questions in the Basic and Clinical Science Course books, which are the main textbooks, are intended to measure the level of learning rather than exam success, such a distinction was not deemed appropriate. Another limitation could have been the comparison of data with human participants. However, it would have been objective to ask and evaluate these questions informally.

As a conclusion, ChatGPT-4.0 gave the highest rate of correct answers to glaucoma questions among LLMs. Bing followed it with a small difference. ChatGPT-3.5 was found to be insufficient. Bard is the most open LLM for improvement even though it has a low rate. ChatGPT-4.0 and Bard are currently the most important candidates as education and training tools in the field of glaucoma. The development of LLMs will strengthen our hand in this field, but misinformation and errors are the most important points to be considered when using these tools.

Author Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Mehmet Canleblebici. The first draft of the manuscript was written by Mehmet Canleblebici, and Ali Dal and Murat Erdağ commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

The authors disclose that they did not receive any funding, grants, or other forms of support while preparing this paper.

Declarations

Conflict of interest

There are no relevant financial or non-financial interests for the authors to state.

Ethical approval

Our study did not require ethics committee approval because it did not include human or animal participants.

REFERENCES

1. Wankhede A, Rajvaidya R, Bagi S. Applications of artificial intelligence and the millennial expectations and outlook towards artificial intelligence. Academy of Marketing Studies Journal. 2021; 25:1-16.
2. Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. arXiv preprint arXiv:230703109. 2023.
3. Meskó B, Hetényi G, Győrffy Z. Will artificial intelligence solve the human resource crisis in healthcare? BMC health services research. 2018; 18:1-4.
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nature medicine. 2023; 29:1930-1940.
5. Safranek CW, Sidamon-Eristoff AE, Gilson A, et al. The role of large language models in medical education: applications and implications. Vol 9: JMIR Publications Toronto, Canada; 2023: e50945.
6. Kung T, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2 (2): e0000198. 2023.
7. Panthier C, Gatinel D. Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: A novel approach to medical knowledge assessment. J Fr Ophtalmol. 2023; 46:706-711.
8. Cai LZ, Shaheen A, Jin A, et al. Performance of Generative Large Language Models on Ophthalmology Board-Style Questions. Am J Ophthalmol. 2023; 254:141-149.
9. Fowler T, Pullen S, Birkett L. Performance of ChatGPT and Bard on the official part 1 FRCOphth practice questions. Br J Ophthalmol. 2023.

10. Sakai D, Maeda T, Ozaki A, et al. Performance of ChatGPT in Board Examinations for Specialists in the Japanese Ophthalmology Society. Cureus. 2023; 15.

11. Li J-PO, Liu H, Ting DS, et al. Digital technology, tele-medicine and artificial intelligence in ophthalmology: A global perspective. Progress in retinal and eye research. 2021; 82:100900.

12. Min B, Ross H, Sulem E, et al. Recent advances in natural language processing via large pre-trained language models: A survey. ACM Computing Surveys. 2023; 56:1-40.

13. Taloni A, Borselli M, Scarsi V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. Sci Rep. 2023; 13:18562.

14. Farhat F, Chaudry BM, Nadeem M, et al. Evaluating AI models for the National Pre-Medical Exam in India: a head-to-head analysis of ChatGPT-3.5, GPT-4 and Bard. JMIR Preprints. 2023.

15. Rudolph J, Tan S, Tan S. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. Journal of Applied Learning and Teaching. 2023; 6.

16. Ahmed I, Roy A, Kajol M, et al. ChatGPT vs. Bard: a comparative study. Authorea Preprints. 2023.

17. McCannel CA, Bhatti MT. The Basic and Clinical Science Course of the American Academy of Ophthalmology: The 50th Anniversary of a Unicorn Among Medical Textbooks. JAMA ophthalmology. 2022; 140:225-226.

18. Sensoy E, Citirik M. A comparative study on the knowledge levels of artificial intelligence programs in diagnosing ophthalmic pathologies and intraocular tumors evaluated their superiority and potential utility. Int Ophthalmol. 2023.

19. Delsoz M, Madadi Y, Munir WM, et al. Performance of ChatGPT in Diagnosis of Corneal Eye Diseases. medRxiv. 2023.

20. Knebel D, Priglinger S, Scherer N, et al. Assessment of ChatGPT in the Prehospital Management of Ophthalmological Emergencies - An Analysis of 10 Fictional Case Vignettes. Klin Monbl Augenheilkd. 2023.

21. Rasmussen MLR, Larsen A-C, Subhi Y, et al. Artificial intelligence-based ChatGPT chatbot responses for patient and parent questions on vernal keratoconjunctivitis. Graefe's Archive for Clinical and Experimental Ophthalmology. 2023; 261:3041-3043.

22. Tan Yip Ming C, Rojas-Carabali W, Cifuentes-González C, et al. The Potential Role of Large Language Models in Uveitis Care: Perspectives After ChatGPT and Bard Launch. Ocul Immunol Inflamm. 2023:1-5.

23. Biswas S, Logan NS, Davies LN, et al. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. Ophthalmic Physiol Opt. 2023; 43:1562-1570.

24. Bernstein IA, Zhang YV, Govil D, et al. Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. JAMA Netw Open. 2023; 6:e2330320.

25. Wu G, Lee DA, Zhao W, et al. ChatGPT: Is it Good for Our Glaucoma Patients? Frontiers in Ophthalmology. 3:1260415.